

Building “Geographic Data Science...”

Dani Arribas-Bel [@darribas]



**The
Alan Turing
Institute**

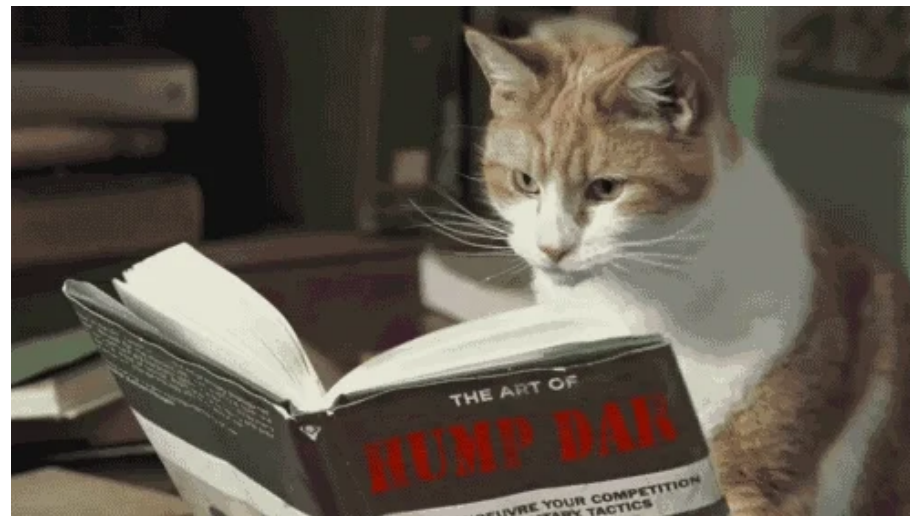


This talk



darribas.org/gdsbook_overview/202109

We have a book!



via GIPHY

Almost...



via GIPHY

Coming “soon” but...

Coming “soon” but...

... you can already:

- <https://geographicdata.science>
- 
- <https://github.com/gdsbook/book>



The Authors



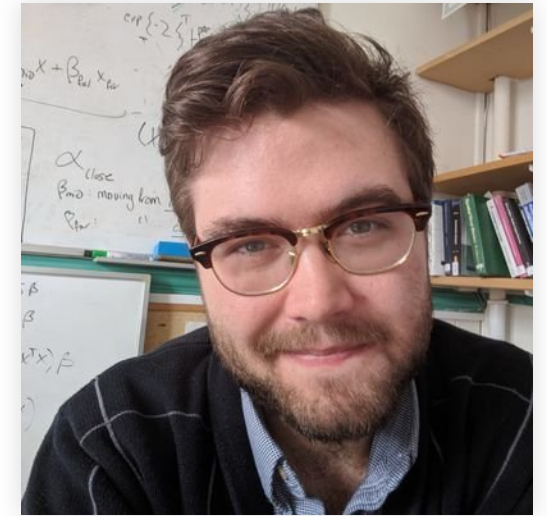
@sreyog

Serge Rey



@darribas

Dani
Arribas-Bel



@levijohnwolf

Levi Wolf

Today

- *Why?* – Principles
- *What?* – Book
- *How?* – Infrastructure & community

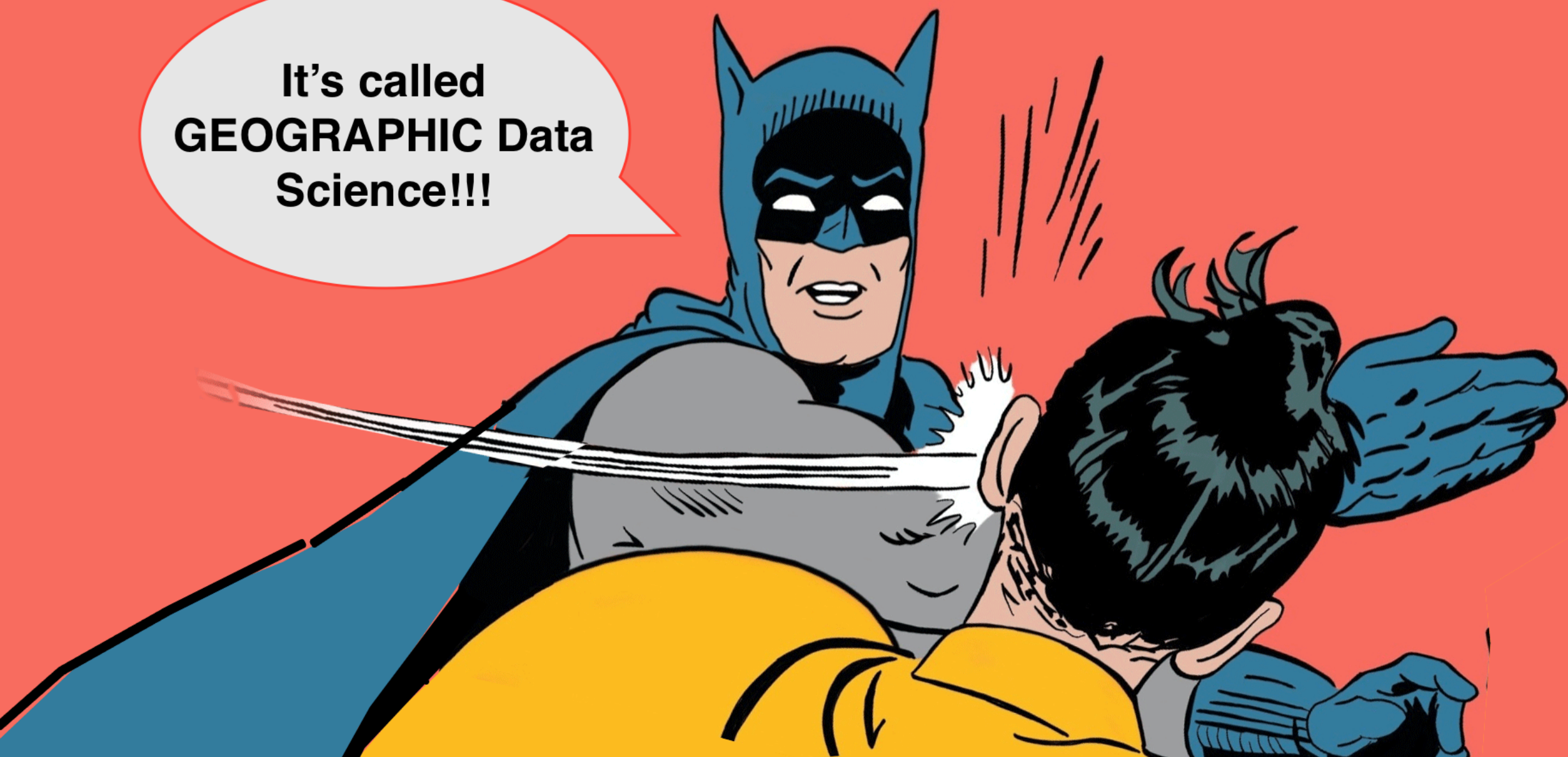
Why?

Data, data, data

Data Science

...

**It's called
GEOGRAPHIC Data
Science!!!**



Geographic Data Science

geographical analysis

Geographical Analysis (2021) **53**, 61–75

Special Issue

Geographic Data Science

Alex Singleton , Daniel Arribas-Bel 

Department of Geography and Planning, University of Liverpool, Liverpool, L69 7ZT, U.K.

It is widely acknowledged that the emergence of “Big Data” is having a profound and often controversial impact on the production of knowledge. In this context, Data Science has developed as an interdisciplinary approach that turns such “Big Data” into information. This article argues for the positive role that Geography can have on Data Science when being applied to spatially explicit problems; and inversely, makes the case that there is much that Geography and Geographical Analysis could learn from Data Science. We propose a deeper integration through an ambitious research agenda, including systems engineering, new methodological development, and work toward addressing some acute challenges around epistemology. We argue that such issues must be resolved in order to realize a Geographic Data Science, and that such goal would be a desirable one.

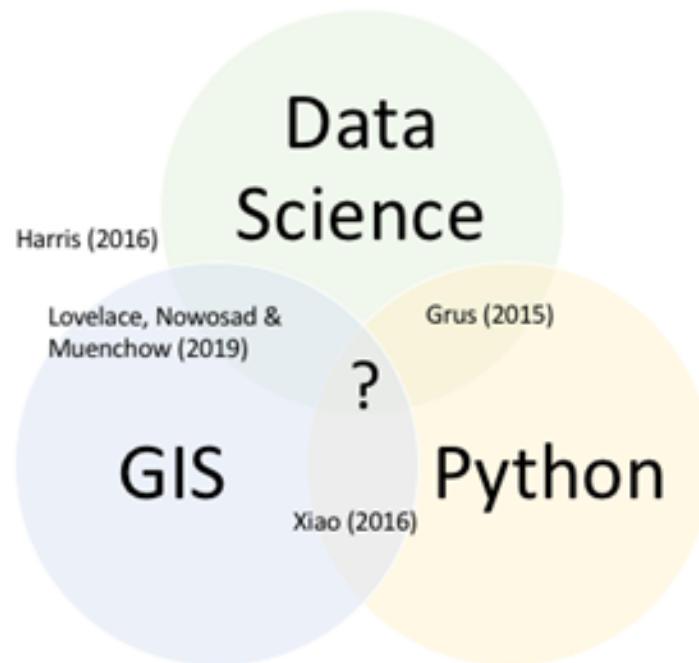


Geographic Data Science

- GIScience ↔ Data Science
- Foster innovation (avoid reinventing the wheel)
- Grow a community around collaboration ($1+1>2$)

What?

The Book is...



The Book is *not*...

- A GIS starter
- An introduction to programming
- An in-depth volume (rather *in-breath*)

The Book is for...

- Data Scientists who work on spatial problems
- GIScientists who want to “update”
- (Social) Scientists getting started in geospatial



Geographic Data Science with Python

Search this book...

[Home](#)

[PREFACE](#)

[Table of Contents](#)

[References](#)

**PART I - BUILDING
BLOCKS**

[Overview](#)

[Geographic thinking for data
scientists](#)

[Computational Tools for
Geographic Data Science](#)

[Spatial Data](#)

[Spatial Weights](#)

**PART II - SPATIAL DATA
ANALYSIS**



Contents

[Part I: Building Blocks](#)

[Part II: Spatial Data Analysis](#)

[Part III: Advanced Topics](#)

Table of Contents

Part I: Building Blocks

- [Geographic Thinking for Data Scientists](#)
- [Geospatial Computational Environment](#)
- [Spatial data](#)
- [Spatial weights](#)

Part II: Spatial Data Analysis

- [Choropleth Mapping](#)
- [Spatial Autocorrelation](#)
- [Local Spatial Autocorrelation](#)
- [Point Pattern Analysis](#)

Part III: Advanced Topics

- [Spatial Inequality](#)
- [Clustering and Regionalization](#)
- [Spatial Regression](#)
- [Spatial Feature Engineering](#)

[previous](#)
[Home](#)

[next](#)
[References](#)

GIScience + Data Science

- geopandas
- rasterio
- pysal
- osmnx
- contextily
- pandas
- xarray
- scikit-learn
- matplotlib
- seaborn

Bonus: Datasets

AirBnb

Download files

Download dataset listing file

```
url = "https://data.icecube.edu/icecube-status/"
headers = {"User-Agent": "icecube-status"}
r = requests.get(url)
with open('icecube-status.txt', 'w') as f:
    f.write(r.text)
```

Download dataset file with prices

```
url = "https://data.icecube.edu/icecube-status/"
headers = {"User-Agent": "icecube-status"}
```

Airports

Download files

Download dataset listing file

```
url = "https://www.naturalsatdata.com/http://www.naturalsatdata.com/download/country/"
headers = {"User-Agent": "icecube-status"}
r = requests.get(url)
with open('country.txt', 'w') as f:
    f.write(r.text)
```

scale	latitude	type	name	abbrev	location	geo_code	lat_code
0	9	Airport	small	Schweid	LLH	terminal	VLD

Brexit

Brexit dataset

This dataset contains results for the Brexit vote at the local authority district, and administrative boundaries.

brexit_vote.csv

Download files

Download dataset listing file

```
url = "https://www.electoralcommission.org.uk/_data/assets/file/014/211310/brexit-vote.csv"
headers = {"User-Agent": "icecube-status"}
```

Countries

Download files

Download dataset listing file

```
url = "https://www.naturalsatdata.com/http://www.naturalsatdata.com/download/country/"
headers = {"User-Agent": "icecube-status"}
r = requests.get(url)
with open('country.txt', 'w') as f:
    f.write(r.text)
```

Download dataset file with prices

```
url = "https://www.naturalsatdata.com/http://www.naturalsatdata.com/download/country/"
headers = {"User-Agent": "icecube-status"}
```

H3 Grid

Build a H3 grid for the San Diego region

Infrastructure

To create a container that includes the following on a file called `h3_grid`

```
docker build -t h3_grid .
```

Mexico

This dataset contains decadal GDP figures for Mexico states, from 1940-2000.

Download files

Download dataset listing file

```
url = "https://data.icecube.edu/icecube-status/"
headers = {"User-Agent": "icecube-status"}
```

Texas

This dataset includes geographic data for Texas counties.

Download files

Download dataset listing file

```
url = "https://www.naturalsatdata.com/http://www.naturalsatdata.com/download/country/"
headers = {"User-Agent": "icecube-status"}
```

Texas

This dataset includes geographic data for Texas counties.

Download files

Download dataset listing file

```
url = "https://www.naturalsatdata.com/http://www.naturalsatdata.com/download/country/"
headers = {"User-Agent": "icecube-status"}
```

Tokyo Photographs

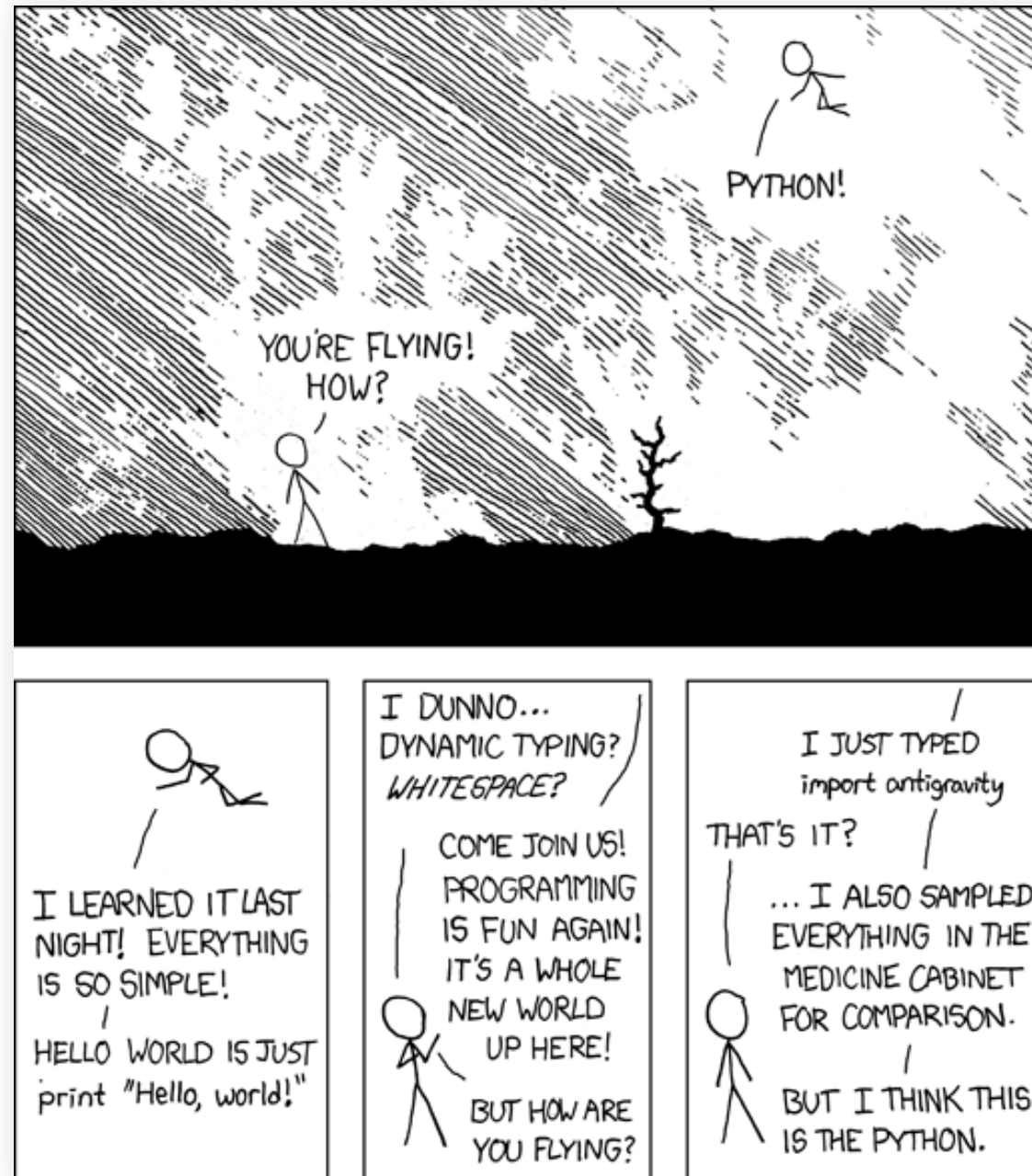
Download files

Download dataset listing file

```
url = "https://www.naturalsatdata.com/http://www.naturalsatdata.com/download/country/"
headers = {"User-Agent": "icecube-status"}
```

How?

Python



Source

Radically Open

Welcome! | Geographic Data Science with PySAL and the PyData Stack

[Authors](#) [Book](#)

Welcome!

Geographic Data Science with PySAL and the PyData Stack

This is the site for the book "Geographic Data Science with PySAL and the PyData Stack", by Sergio J. Rey, Dani Arribas-Bel and Levi J. Wolf. Here you can find out about the latest news regarding the book, read more about the [authors](#), or jump straight to the [book](#).

Latest news

Oct 30, 2020
[Geographic data science on the road!](#)

Apr 21, 2020
[New Data Section available online](#)

Aug 29, 2019
[Work in progress...](#)

Aug 24, 2019
[Hello world!!!](#)

[Subscribe](#)

Geographic Data Science with PySAL and the PyData Stack

Sergio J. Rey, Dani Arribas-Bel & Levi J. Wolf
geographicdatascience@gmail.com

This book serves as an introduction to a whole new way of thinking systematically about geographic data, using geographical analysis and computation to unlock new insights hidden within data.

[🔄](#) [🐦](#) [📡](#)

GitHub - gdsbook/book: This book serves as an introduction to a whole new way of thinking systematically about geographic data, using geographical analysis and computation to unlock new insights hidden within data.

gdsbook / book

Code Issues 23 Pull requests 5 Actions Projects Security Insights

master

actions-user GA build of book HTML 12 days ago 341

.github	Change GA commit message to clarify further	2 months ago
data	Fix data downloads and update gds_env for ...	2 months ago
docs	GA build of book HTML	12 days ago
figures	Complete draft of ch01	11 months ago
infrastructure	Fix titles so they look OK on side TOC	2 months ago
notebooks	GA build of book HTML	12 days ago
.gitignore	Ignore NASADEM .hgt files	8 months ago
.nojekyll	Create .nojekyll	2 years ago
Dockerfile	Removed paired markdowns on Binder setup	2 months ago
LICENSE	restructure of infrastructure to move to new JB	2 months ago
Makefile	Remove CNAME (taken care of in parent rep...	2 months ago
README.md	Update README.md	15 months ago
appveyor.yml	add substance and infrastructure	2 years ago

About

This book serves as an introduction to a whole new way of thinking systematically about geographic data, using geographical analysis and computation to unlock new insights hidden within data.

[geographicdata.science](#)

- [data-science](#)
- [data-analysis-python](#)
- [geographical-information-system](#)
- [geographic-data](#)
- [spatial-analysis](#)
- [spatial-statistics](#)
- [statistics](#)
- [spatial-data-analysis](#)

Readme

View license

Commits - gdsbook/book · GitHub

gdsbook / book

Code Issues 23 Pull requests 5 Actions Projects Security Insights

master

Commits on Nov 6, 2020

- GA build of book HTML actions-user committed 12 days ago ✓ 7375dca
- Merge pull request #102 from ljwolf/ch2 Verified darribas committed 12 days ago ✓ 4762249
- update spatial data chapter with additional structure ljwolf committed 12 days ago 63adeb9

Commits on Oct 21, 2020

- GA build of book HTML actions-user committed 29 days ago ✓ d8631b1
- Merge pull request #100 from darribas/master Verified ljwolf committed 29 days ago ✓ d24db1a
- GA build of book HTML actions-user committed 29 days ago 0400d49
- Ch.12: Remove DEM code to go to blog: fix header hierarchy for cluste... darribas committed 29 days ago 6bbfacb

Commits on Oct 16, 2020

- add sketch of spatial data chapter ljwolf committed on Oct 16 e67db76

Commits on Sep 21, 2020

- GA build of book HTML actions-user committed on Sep 21 310ead1
- Add refs page darribas committed on Sep 21 39d1da4
- interim build

Code as text; text as code

Geographic Data Science with Python

Search this book...

Home

PREFACE

Table of Contents

References

PART I - BUILDING BLOCKS

Overview

Geospatial Computational Environment

Geographic thinking for data scientists

Spatial Data Processing

Spatial Weights

PART II - SPATIAL DATA ANALYSIS

Overview

Choropleth Mapping

Global Spatial Autocorrelation

Local Spatial Autocorrelation

Point Pattern Analysis

PART III - ADVANCED TOPICS

Overview

Spatial Inequality

Clustering & Regionalization

Spatial Regression

Spatial Feature Engineering

DATASETS

Overview

AirBnb

Airports

Brexit

Countries

H3 Grid

Mexico

Contents

Introduction

Location, Location, Location

Visualization

Centrality

Randomness & clustering

Identifying clusters

Conclusion

Questions

References

To illustrate, the figure below has the tightest single alpha shape shown in green and the original source points shown in black. The "bounding" circles shown in the figure all have a radius of 8652 meters. The circles are plotted where our "bounding" disk touches two or three of the points in the point cloud. You can see that the circles "cut into" the convex hull, shown in blue dashed lines, up until they touch two (or three) points. Any tighter, and the circle would disconnect one of the points on the boundary of the alpha shape.

```
from descartes import PolygonPatch #to plot the alpha shape easily
f,ax = plt.subplots(1,1, figsize=(9,9))

# Plot a green alpha shape
ax.add_patch(PolygonPatch(alpha_shape, edgecolor='green',
                           facecolor='green', alpha=.2, label = 'Tightest alpha shape'))

# Include the points for our prolific user in black
ax.scatter(*coordinates.T, color='k', marker='.', label='Source Points')
# Add a basemap
ax.imshow(basemap, extent=basemap_extent)

# plot the circles forming the boundary of the alpha shape
for i, circle in enumerate(circles):
    # only label the first circle of its kind
    if i == 0:
        label = 'Bounding Circles'
    else:
        label = None
    ax.add_patch(plt.Circle(circle, radius=alpha, facecolor='none', edgecolor='r', label=label))

# add a blue convex hull
ax.add_patch(plt.Polygon(convex_hull_vertices,
                        closed=True,
                        edgecolor='blue', facecolor='none',
                        linestyle='--', linewidth=2,
                        label='Convex Hull'))

plt.legend()
```

<matplotlib.legend.Legend at 0x7f413c088c10>

Merge pull request #100 from darribas/master

https://github.com/gdsbook/book/runs/1285952518?ci=80%

Why GitHub? Team Enterprise Explore Marketplace Pricing

Search Sign in Sign up

gdsbook / book

Watch 11 Star 108 Fork 25

Code Issues 23 Pull requests 5 Actions Projects Security Insights

Merge pull request #100 from darribas/master

master d24db1a

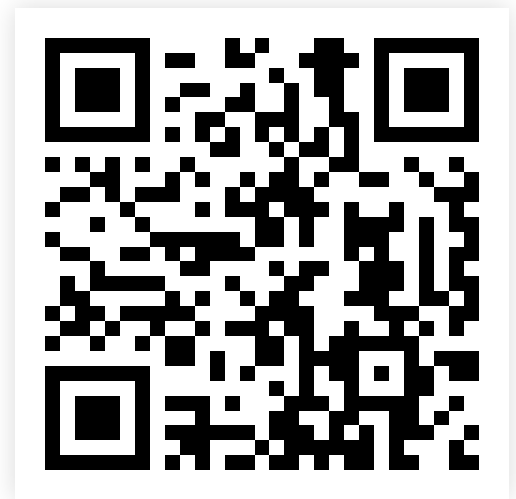
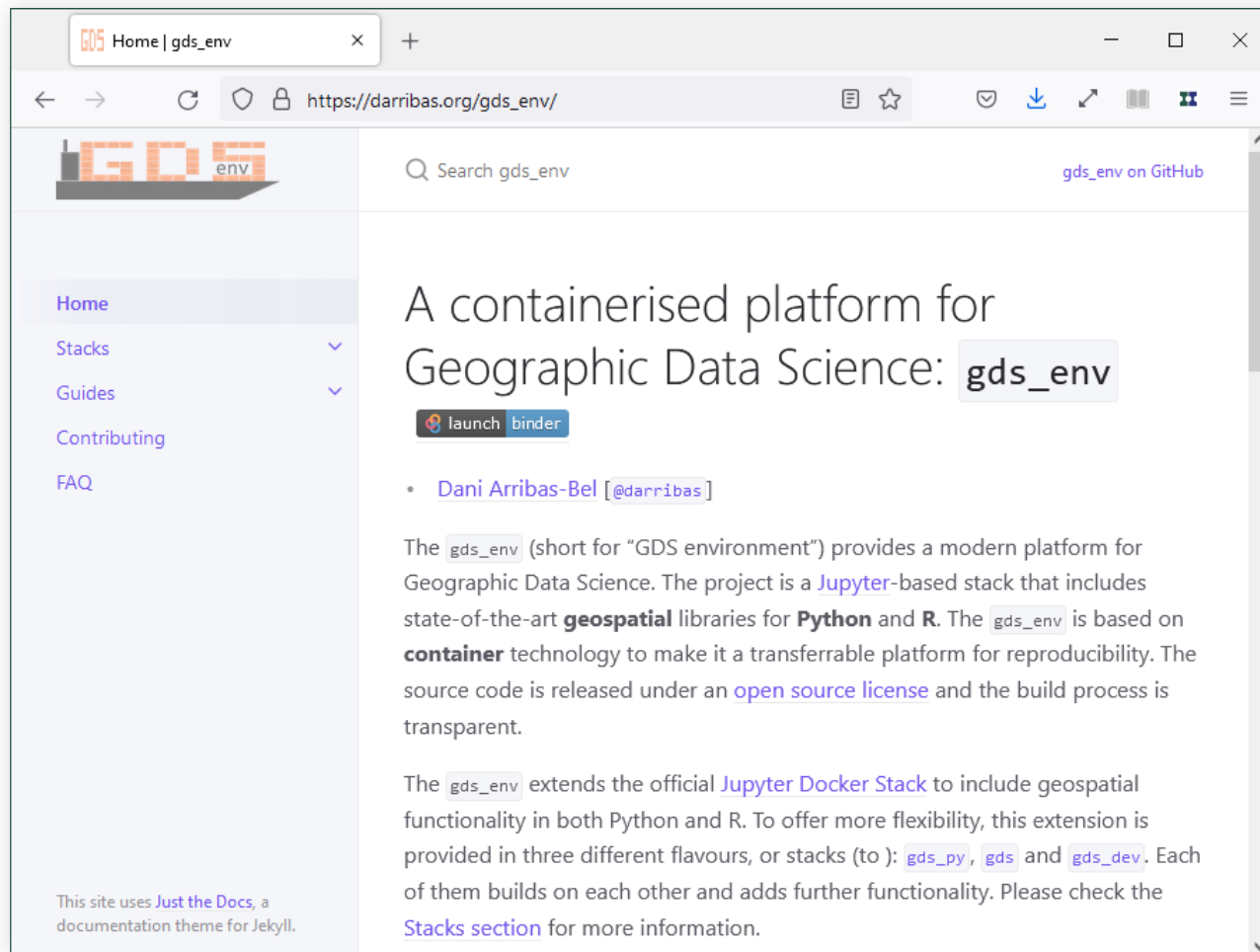
Build Jupyter book on: push 6

build-html-and-deploy

succeeded 29 days ago in 2m 50s

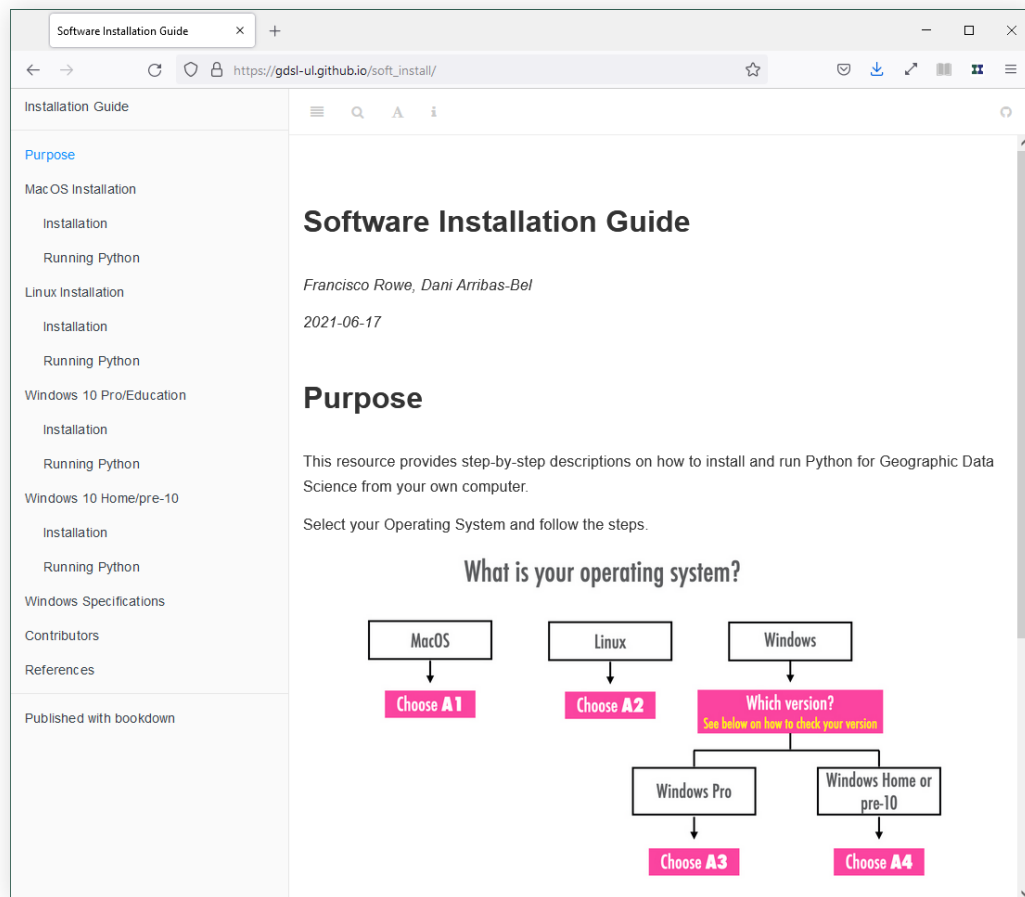
- Set up job 4s
- Checkout 13s
- Setup Miniconda 1m 59s
- Build HTML 29s
- Commit files 0s
- Push changes 4s
- Post Setup Miniconda 1s
- Post Checkout 0s
- Complete job 0s

Runs *anywhere*...



darribas.org/gds_env

... by anyone



`gds1-ul.github.io/soft_install`

Try it out...

Spatial Feature Engineering — X +

← → ↻ <https://geographicdata.science> 90% ... ☆

← ↻ ↗


Geographic Data Science with Python

🔍 Search this book

Home

PREFACE

Table of Contents

References

Spatial Feature Engineering

Launch Binder

 Binder

 Colab

In machine learning and data science, we are often equipped with *tons* of data. Indeed, given the constellation of packages to query data services, free and open source data sets, and the rapid and persistent collection of geographical data, there is simply too much data to even represent coherently in a single, tidy fashion. However, we often need to be able to construct useful *features* from this rich and deep sea of data.

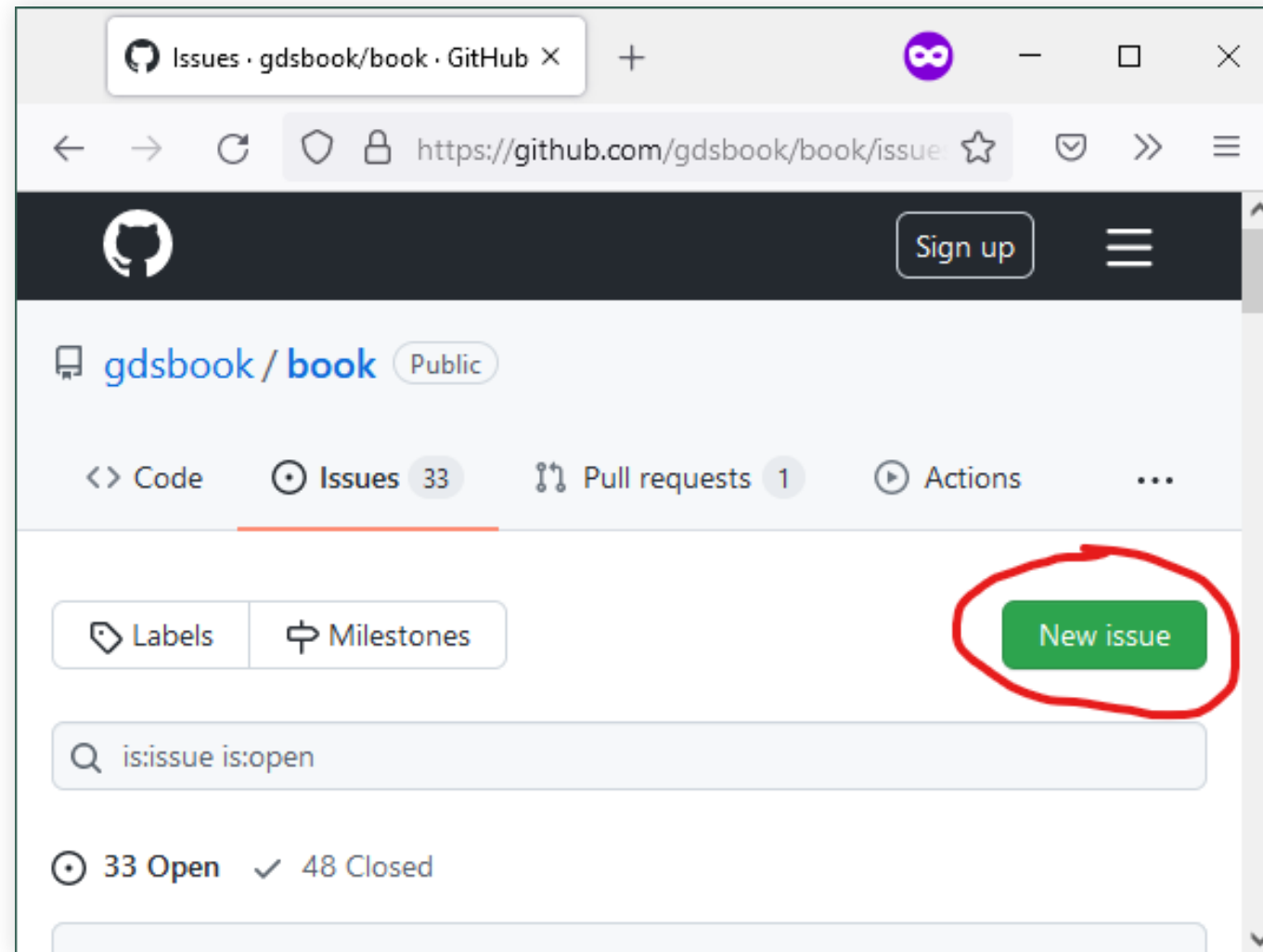
Where data is available, but not yet directly *usable*, *feature engineering* helps to construct useful data for modelling a

Contents

- What is spatial feature engineering?
- Feature Engineering Using Map Matching
- Feature Engineering using Map Synthesis
- Conclusion
- Questions

https://mybinder.org/v2/gh/gdsbook/book/master?urlpath=lab/tree/notebooks/12_feature_engineering.ipynb

...and make it better!!!



`github.com/gdsbook/book/issues/new`

Building “Geographic Data Science...”

Dani Arribas-Bel [@darribas]



UNIVERSITY OF
LIVERPOOL

**The
Alan Turing
Institute**



Geographic
Data Science
Lab

[PDF version of these slides]



These slides are licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.